Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 2.1: Basic Concepts – Neural Networks

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.*
Spring 2021 edition taught by Yonatan Bisk

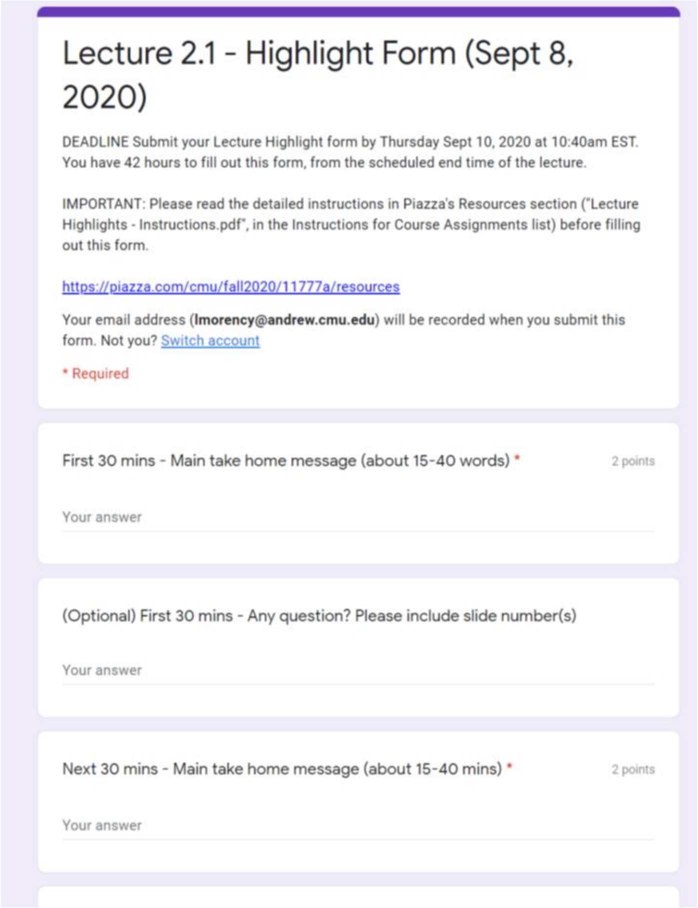
Lecture Objectives

- Unimodal basic representations
  - Visual, language and acoustic modalities
- Data-driven machine learning
  - Training, validation and testing
  - Example: K-nearest neighbor
- Linear Classification
  - Score function
  - Two loss functions (cross-entropy and hinge loss)
- Neural networks

Administrative Stuff

Lecture Highlight Form

https://forms.gle/u3JuHoQhhUDRG3KY7



Lecture 2.1 - Highlight Form (Sept 8, 2020)

DEADLINE Submit your Lecture Highlight form by Thursday Sept 10, 2020 at 10:40am EST. You have 42 hours to fill out this form, from the scheduled end time of the lecture.

IMPORTANT: Please read the detailed instructions in Piazza's Resources section ("Lecture Highlights - Instructions.pdf", in the Instructions for Course Assignments list) before filling out this form.

https://piazza.com/cmu/fall2020/11777a/resources

Your email address (**lmorency@andrew.cmu.edu**) will be recorded when you submit this form. Not you? Switch account

* Required

First 30 mins - Main take home message (about 15-40 words) * — 2 points

Your answer

(Optional) First 30 mins - Any question? Please include slide number(s)

Your answer

Next 30 mins - Main take home message (about 15-40 mins) * — 2 points

Your answer

Deadline: Tuesday 11:59pm ET

(for Thursday's lecture, the deadline is Thursday 11:59pm ET)

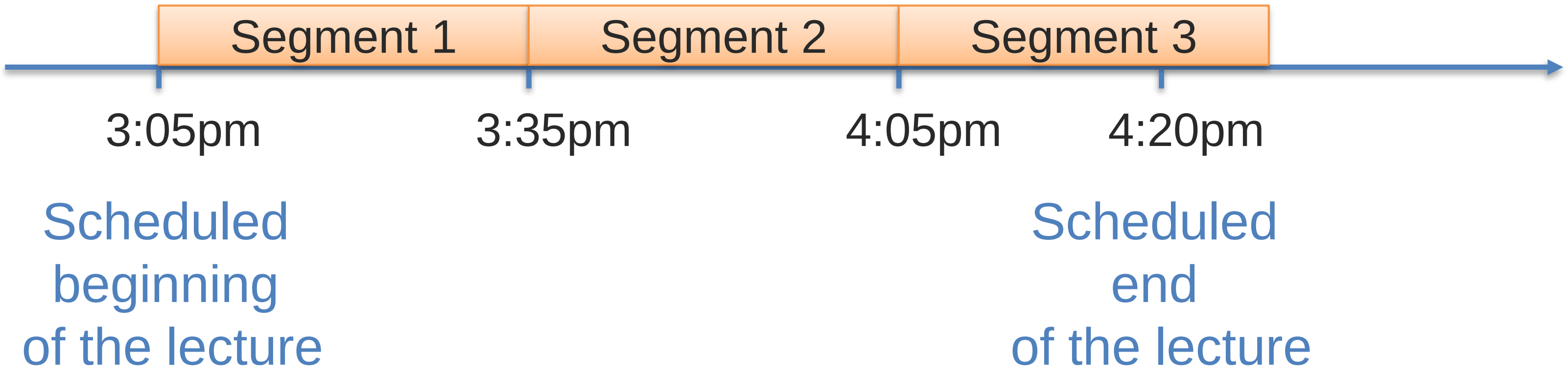
Use your Andrew CMU email

- You will need to login using this address

New form for each lecture

- Posted on Piazza's Resources section

Lecture Highlight Form - Segments



- Segment 1 starts at 3:05pm, even if the lecture starts slightly later.
- Segment 3 ends whenever the lecture ends
- Slides happening around the segment borders (+/- 5min of 3:35pm and 4:05pm) can be included in either neighboring segment.

Lecture Highlight Form - Grading

For each segment

- Two sentences (10+ words each; complete English sentences) describing the two main points described in this segment

For the whole lecture

- Your main two take-aways from the lecture
- About 15-40 words per take-home message
  - Try to be succinct, but with complete English sentences.
- Be concrete in your take-home messages
  - Avoid generic summaries like: “This is about multimodal”

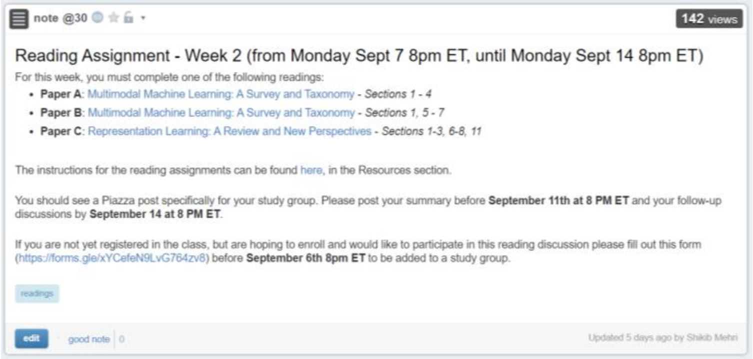
Each submission is worth 1 point

- Final grade is the sum of your top 15 submissions

Reading Assignments – Piazza Posts

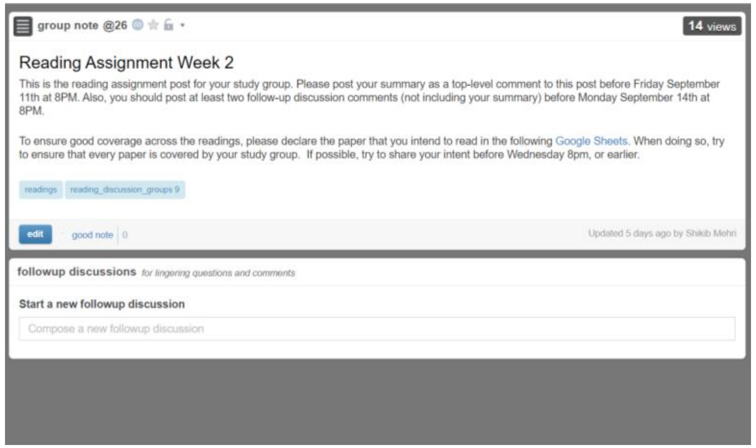
For each reading assignment, 2 instruction posts will be created:

1



- Sent to everyone
- Contains list of reading options

2

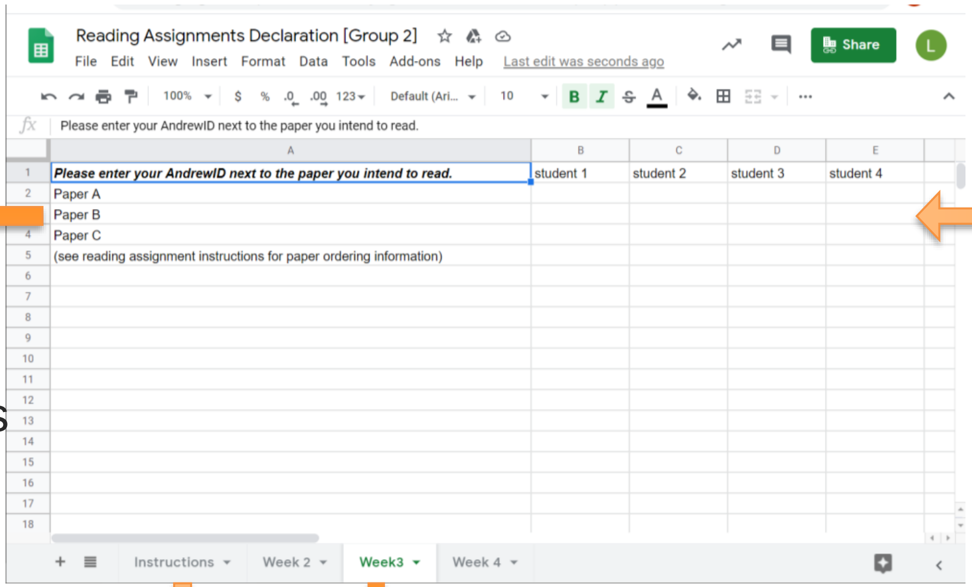


- Sent separately to each study group
- Link to personalized signup sheet
- Post your summary as top-level
- Post your follow-up posts

Reading Assignments – Signup Sheet

Each study group has its own signup sheet:

Sign-up here for the paper option you would like to read and summarize



The details for the paper options are in the first Piazza post

It also contains the list of members in this study group

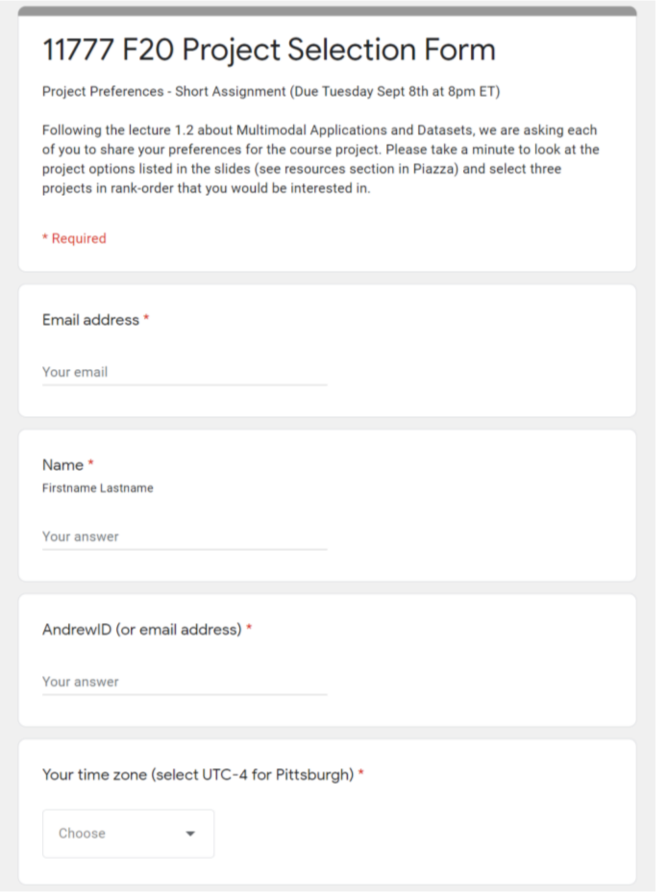
A different tab for each reading assignment

Reading Assignments – Weekly Schedule

Four main steps for the reading assignments

1. **Monday 8pm:** Official start of the assignment
2. **Wednesday 8pm:** Select your paper
3. **Friday 8pm:** Post your summary
4. **Monday 8pm:** End of the reading assignment

Team Matching – Project Preference Form



11777 F20 Project Selection Form

Project Preferences - Short Assignment (Due Tuesday Sept 8th at 8pm ET)

Following the lecture 1.2 about Multimodal Applications and Datasets, we are asking each of you to share your preferences for the course project. Please take a minute to look at the project options listed in the slides (see resources section in Piazza) and select three projects in rank-order that you would be interested in.

* Required

Email address *

Your email

Name *

Firstname Lastname

Your answer

AndrewID (or email address) *

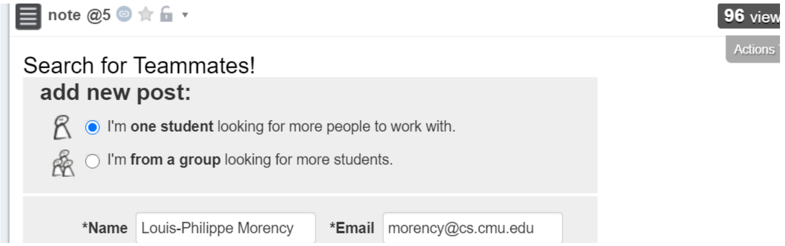
Your answer

Your time zone (select UTC-4 for Pittsburgh) *

Choose

Deadline: Today at 8pm!!

- Every students should submit a form
- Students on the waitlist are also encouraged to submit a form
- A summary will be shared to help you find potential teammates
- Also, you can use Piazza to share info and contact potential teammates



Team Matching – Thursday Event

Thursday around 3:50pm ET

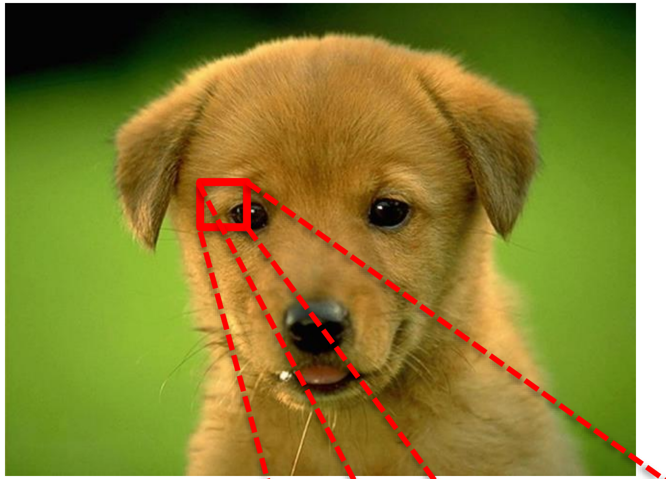
(later part of the lecture)

- Detailed instructions will be shared during lecture
- Event optional for students who already have a full team

Unimodal Basic Representations

Unimodal Representation – Visual Modality

Color image



Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation x_i

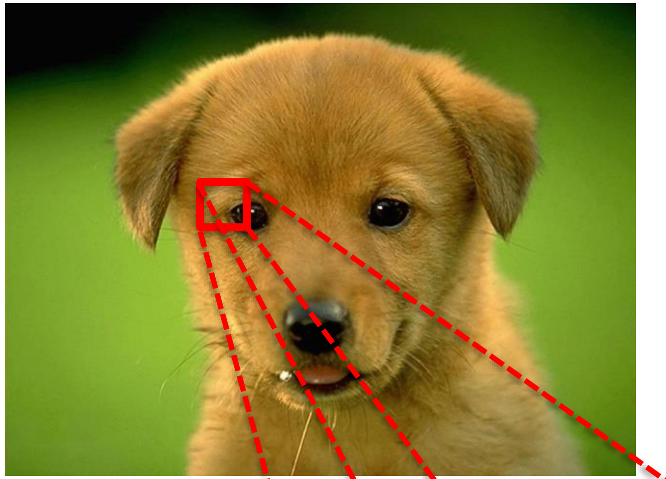
88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Binary classification problem

Dog ?

label $y_i \in \mathcal{Y} = \{0,1\}$

Unimodal Representation – Visual Modality



Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation $\boldsymbol{x_i}$

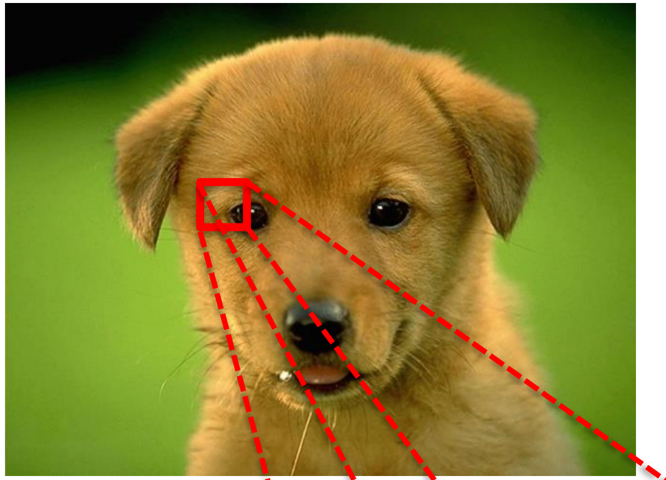
88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-class classification problem

Duck
-or-
Cat
-or-
Dog
-or-
Pig
-or-
Bird ?

label $y_i \in \mathcal{Y} = \{0,1,2,3, \ldots\}$

Unimodal Representation – Visual Modality



Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation $\boldsymbol{x}_i$

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-label (or multi-task) classification problem

Duck?

Cat ?

Dog ?

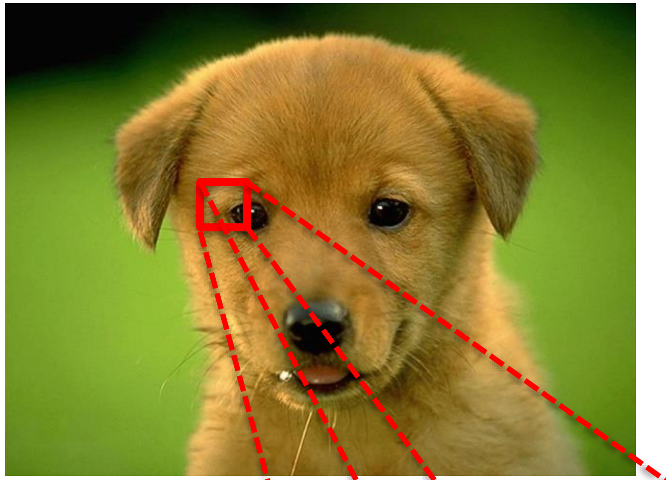
Pig ?

Bird ?

Puppy ?

label vector $\boldsymbol{y}_i \in \mathcal{Y}^m = \{0,1\}^m$

Unimodal Representation – Visual Modality



Each pixel is represented in $\mathbb{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation $\boldsymbol{x}_i$

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-label (or multi-task) regression problem

Age ?

Height ?

Weight ?

Distance ?

Happy ?

label vector $\boldsymbol{y}_i \in \mathcal{Y}^m = \mathbb{R}^m$

Unimodal Representation – Language Modality

Written language

★★★★★ **Masterful!**

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

```
                MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

                CLARK
But you've only put up the good
things they say about me.

                MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation x_i

0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 ⋮

"one-hot" vector

$|x_i|$ = number of words in dictionary

Word-level classification

Part-of-speech ?
(noun, verb,…)

Sentiment ?
(positive or negative)

Named entity ?
(names of person,…)

Unimodal Representation – Language Modality

Written language

Masterful!

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

```
                MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

                CLARK
But you've only put up the good
things they say about me.

                MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation x_i

0
1
0
0
1
0
1
0
0
0
0
1
0
0
0
0
1
0
0
0
⋮

“bag-of-word” vector

$|x_i|$ = number of words in dictionary

Document-level classification

Sentiment ?

(positive or negative)

Unimodal Representation – Language Modality

Written language

★★★★★ **Masterful!**

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

```
              MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

              CLARK
But you've only put up the good
things they say about me.

              MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation x_i

0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, ⋮

"bag-of-word" vector

$|x_i|$ = number of words in dictionary

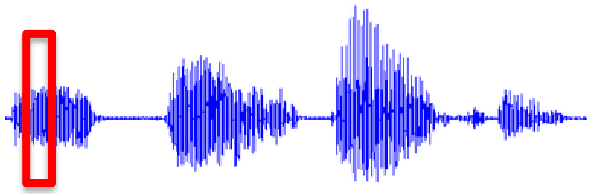
Utterance-level classification

Sentiment ?

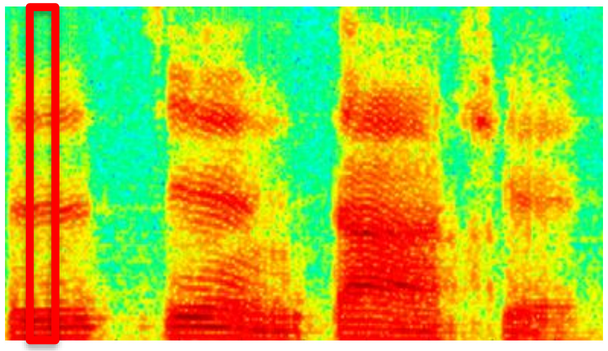
(positive or negative)

Unimodal Representation – Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectogram

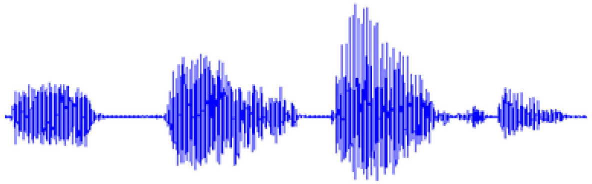
Input observation x_i

0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02

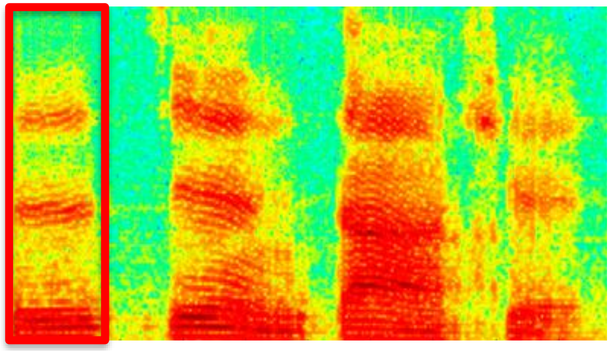
Spoken word ?

Unimodal Representation – Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectogram

Input observation x_i

0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02
0.24
0.26
0.58
0.9
0.99
0.79
0.45
0.34
0.24
⋮

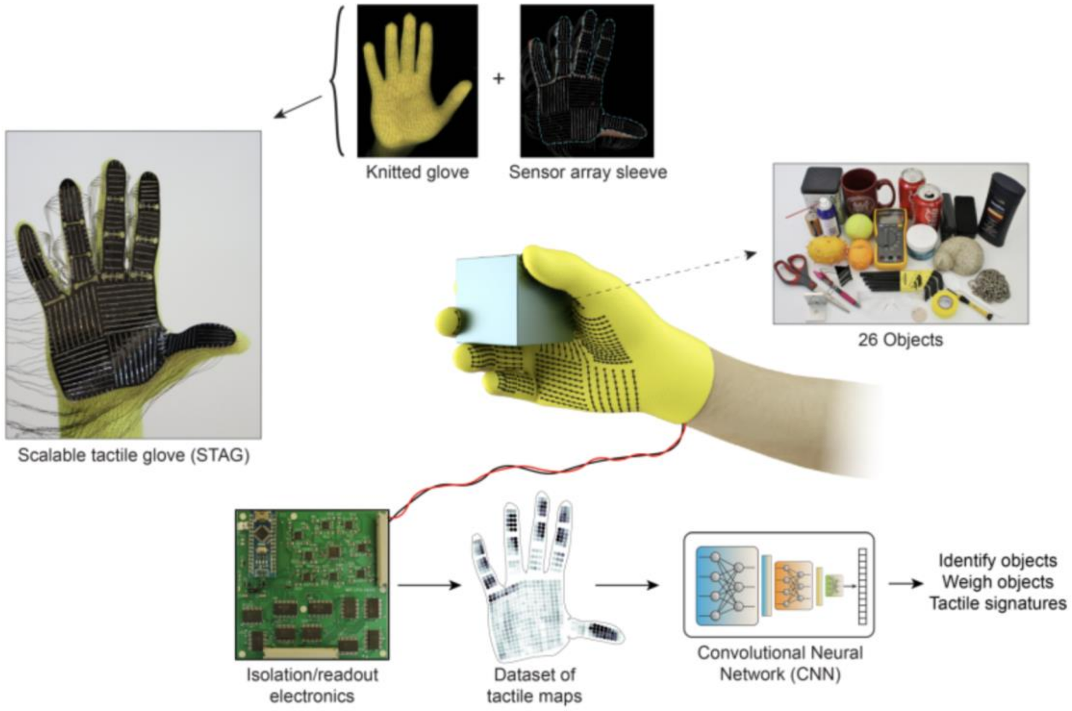
Emotion ?

Spoken word ?

Voice quality ?

Other Unimodal Representations

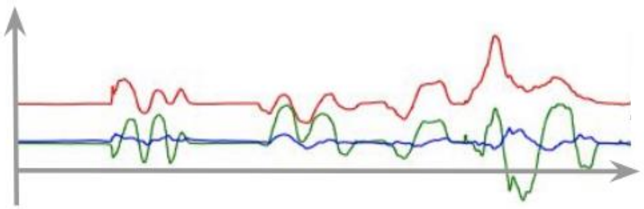
Unimodal Representation – Sensors



The tactile sensor array (548 sensors) is assembled on a knitted glove uniformly distributed over the hand.

Sundaram et al., Learning the signatures of the human grasp using a scalable tactile glove. Nature 2019

Unimodal Representation – Sensors



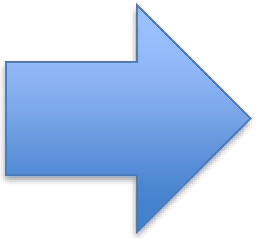
Force-Torque Sensor

Time series data across six-axis Force-Torque sensor:
T × 6 signal.

Proprioception

Measure values internal to the system (robot); e.g. motor speed, wheel load, **robot arm joint angles**, battery voltage.

Time series data across current position and velocity of the end-effector:
T × 2d signal.



Next action

Lee et al., Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. ICRA 2019

Unimodal Representation – Tables

Caption: Singapore Cup

Attribute / Cell:

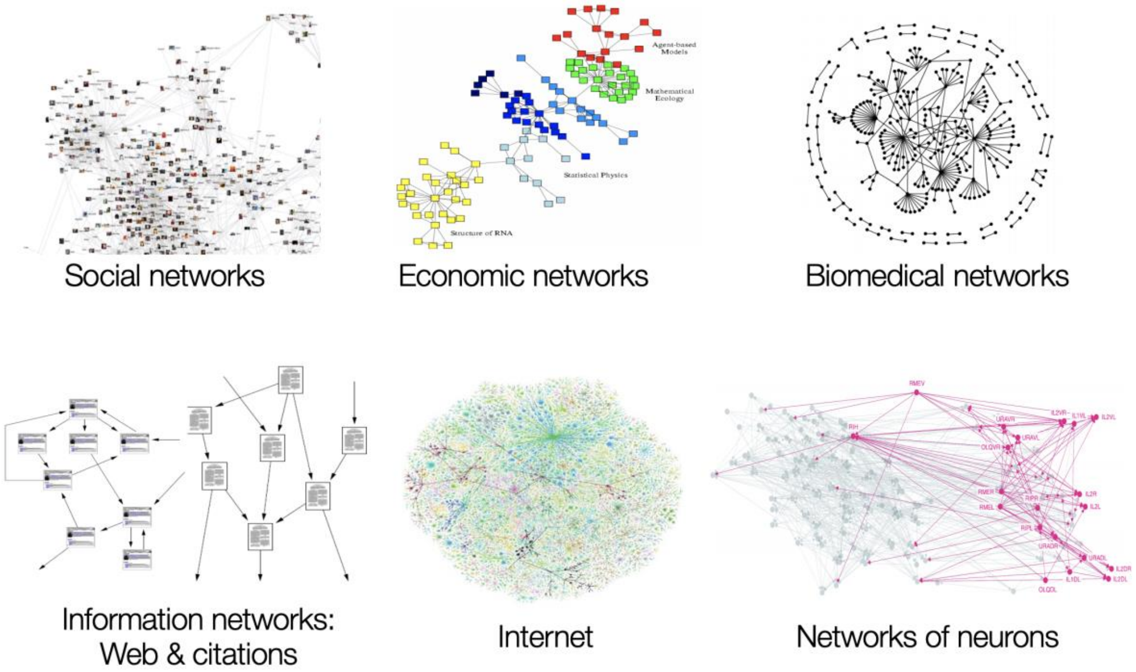
Year	Champions	Runners-up
1996	Geylang United	Singapore Armed Forces
1997	Singapore Armed Forces	Woodlands Wellington
1998	Tanjong Pagar United	Sembawang Rangers

Text: *Singapore Armed forces was the champion of Singapore Cup in 1997.*

Table-to-text generation

Bao et al., Table-to-Text: Describing Table Region with Natural Language. AAAI 2018

Unimodal Representation – Graphs



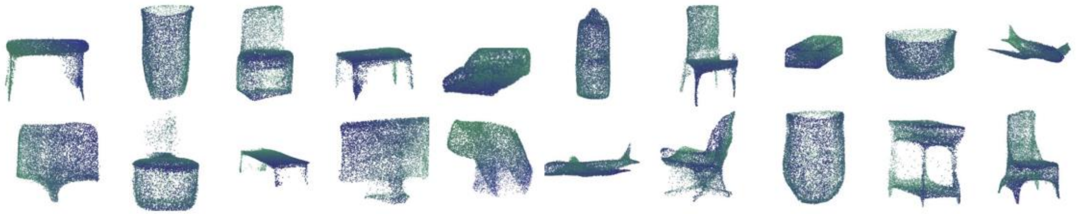
Tasks on graphs:
Node classification
Link prediction
…

Using graphs:
Knowledge graphs
for QA
Social network for
sentiment analysis
…

Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019

Unimodal Representation – Sets

Sets



Point clouds

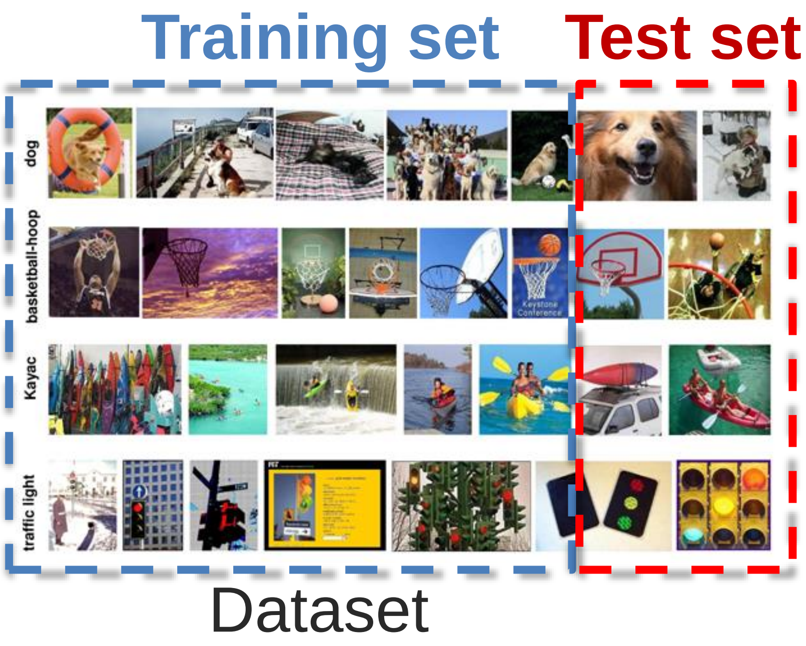
Set anomaly detection
Set expansion
Set completion
Point cloud classification
Point cloud generation

Zaheer et al., DeepSets. NeurIPS 2017, Li et al., Point Cloud GAN. arxiv 2018

Machine Learning – Basic Concepts

Training, Testing and Dataset

1. **Dataset:** Collection of labeled samples $D: \{x_i, y_i\}$
2. **Training:** Learn classifier on training set
3. **Testing:** Evaluate classifier on hold-out test set



Simple Classifier ?

Traffic light

-or-

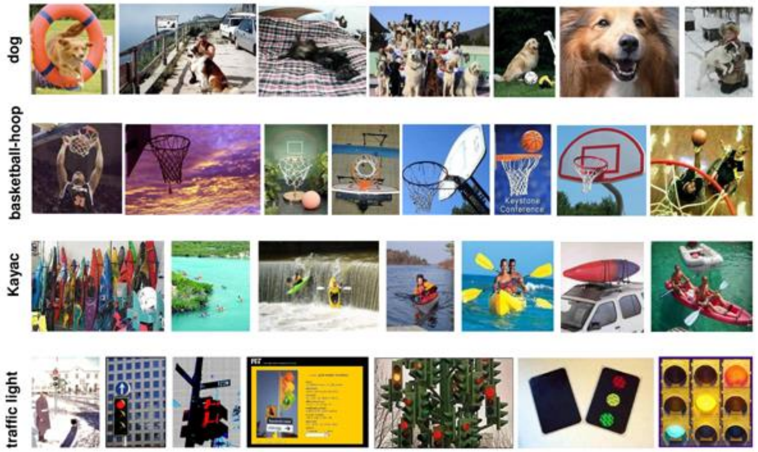
Dog

-or-

Basket

-or-

Kayak ?



Dataset

Simple Classifier: Nearest Neighbor

Traffic light

-or-

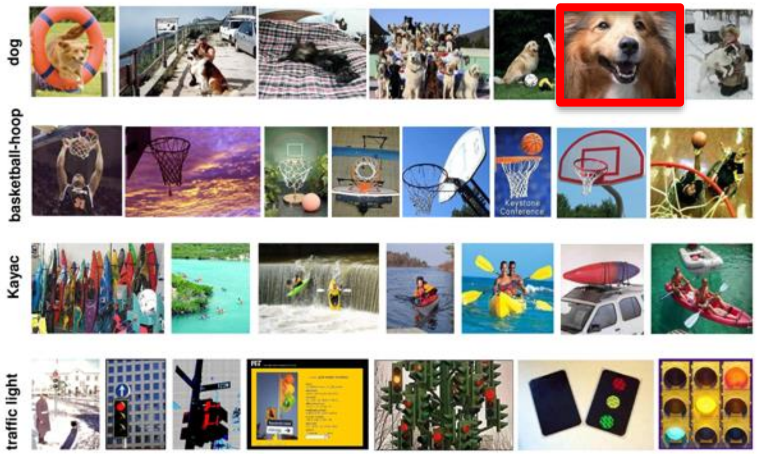
Dog

-or-

Basket

-or-

Kayak ?



Training

Nearest Neighbor Classifier

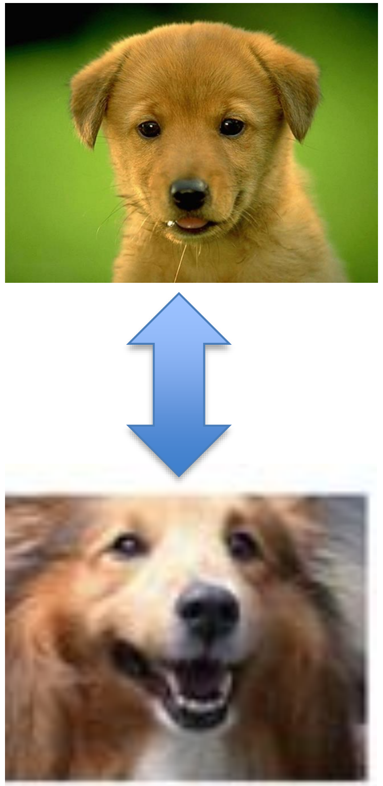
- Non-parametric approaches—key ideas:
  - *“Let the data speak for themselves”*
  - *“Predict new cases based on similar cases”*
  - *“Use multiple local models instead of a single global model”*

- What is the complexity of the NN classifier w.r.t training set of N images and test set of M images?
  - at training time?

    O(1)
  - At test time?

    O(N)

Simple Classifier: Nearest Neighbor



Distance metrics

L1 (Manhattan) distance:

$$d_1(x_1, x_2) = \sum_j \left| x_1^j - x_2^j \right|$$

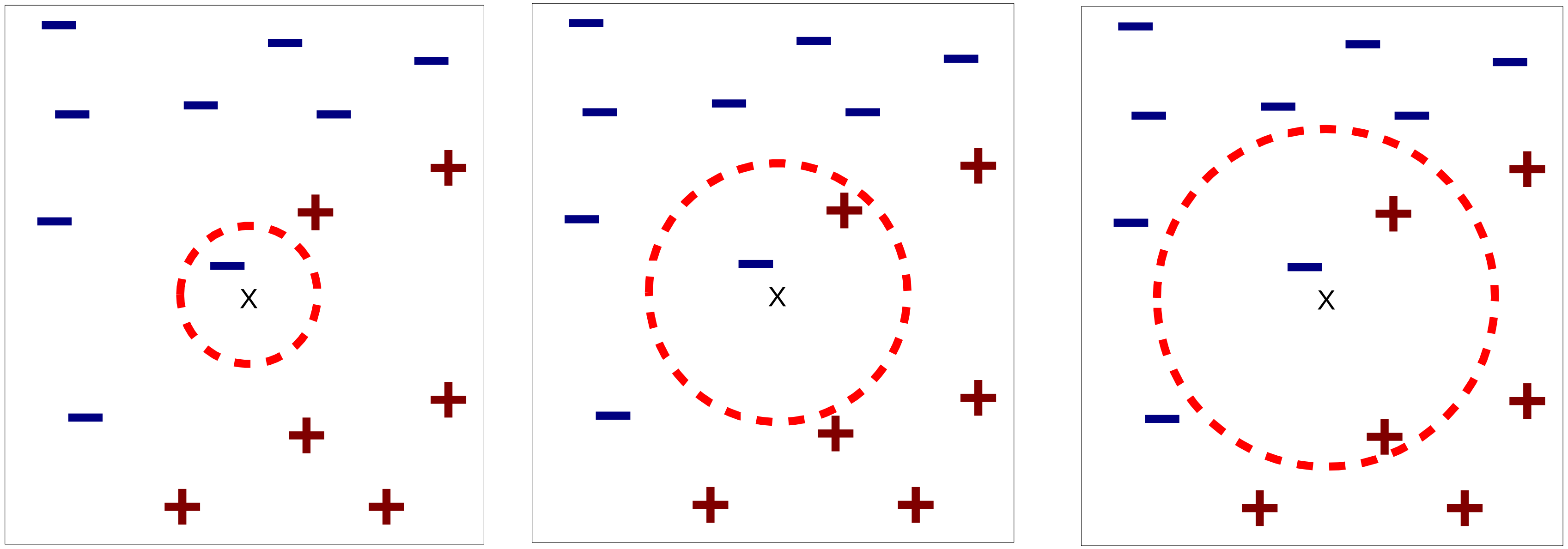
L2 (Eucledian) distance:

$$d_2(x_1, x_2) = \sqrt{\sum_j \left(x_1^j - x_2^j \right)^2}$$

Which distance metric to use?

First hyper-parameter!

Definition of K-Nearest Neighbor



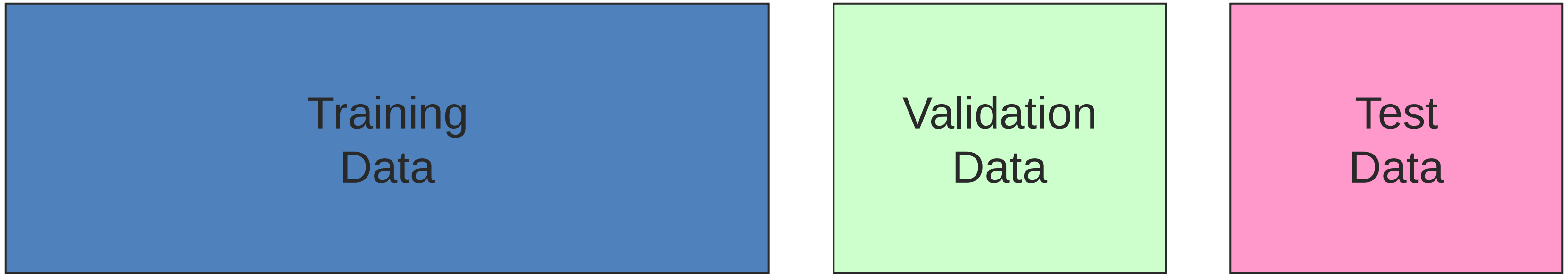
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

What value should we set K?

Second hyper-parameter!

Data-Driven Approach

1. **Dataset:** Collection of labeled samples $\mathrm{D}: \{x_i, y_i\}$
2. **Training:** Learn classifier on training set
3. **Validation:** Select optimal hyper-parameters
4. **Testing:** Evaluate classifier on hold-out test set



Evaluation methods (for validation and testing)

- Holdout set: The available data set *D* is divided into two disjoint subsets,
  - the *training set* D_{train} (for learning a model)
  - the *test set* D_{test} (for testing the model)

- n-fold cross-validation: The available data is partitioned into *n* equal-size disjoint subsets.

- Leave-one-out cross-validation: This method is used when the data set is very small.

Linear Classification: Scores and Loss

Linear Classification (e.g., neural network)

Image

(Size: 32*32*3)

?

1. **Define a (linear) score function**
2. **Define the loss function (possibly nonlinear)**
3. **Optimization**

1) Score Function

Image

(Size: 32*32*3)

Duck ?
Cat ?
Dog ?
Pig ?
Bird ?

What should be the prediction score for each label class?

For linear classifier:

Input observation (i^{th} *element of the dataset*) [3072x1]

$$f(x_i; W, b) = W x_i + b$$

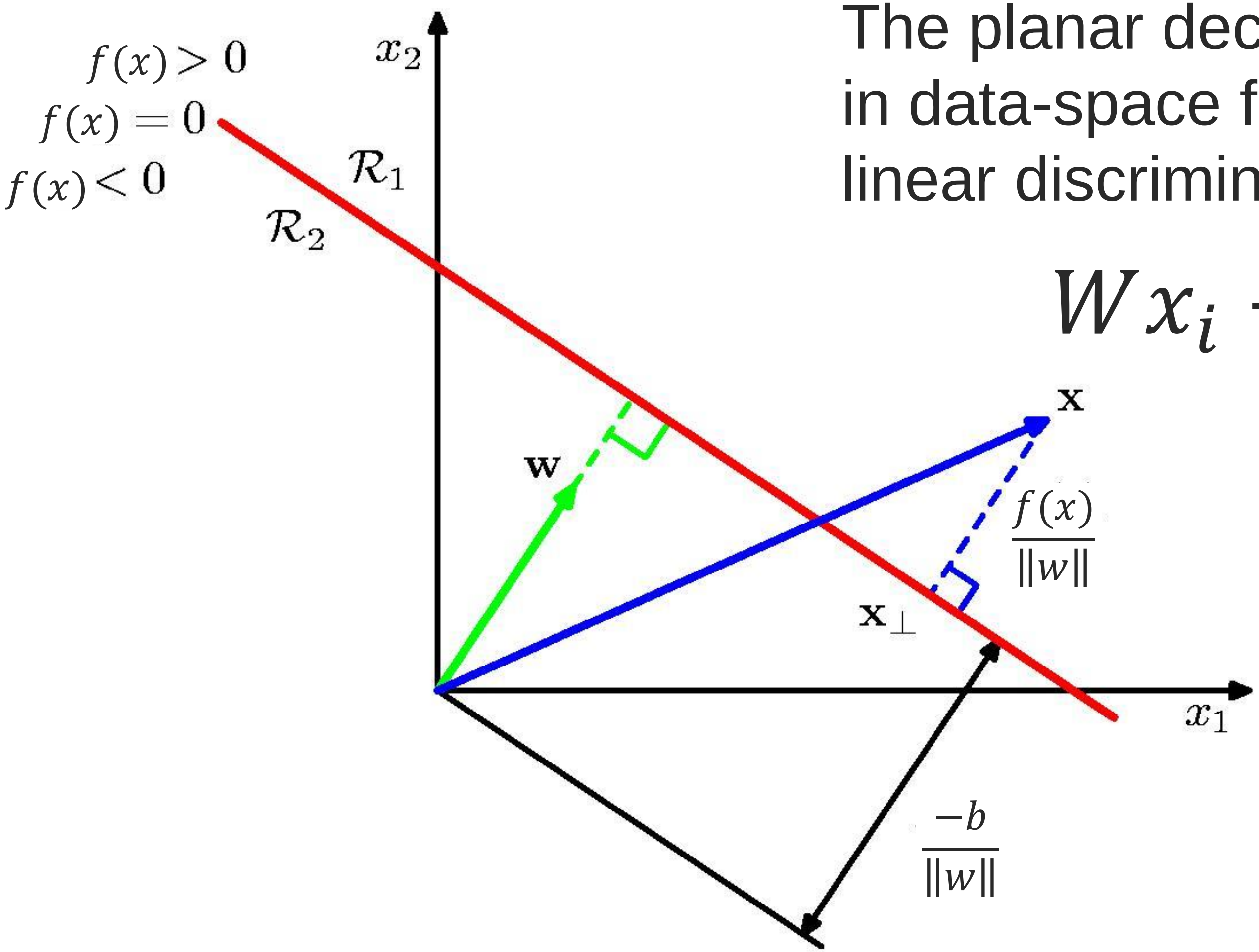
Class score [10x1]

Weights [10x3072]

Bias vector [10x1]

Parameters [10x3073]

Interpreting a Linear Classifier



The planar decision surface in data-space for the simple linear discriminant function:

$$Wx_i + b > 0$$

Some Notation Tricks – Multi-Label Classification

$$W = [W_1 \quad W_2 \quad \dots \quad W_N]$$

$$f(x_i; W, b) = Wx_i + b \quad \rightarrow \quad f(x_i; W) = Wx_i$$

Weights x Input + Bias

[10x3072] [3072x1] [10x1]

Weights x Input

[10x3073] [3073x1]

The bias vector will be the last column of the weight matrix

Add a “1” at the end of the input observation vector

Some Notation Tricks

General formulation of linear classifier: $f(x_i; W, b)$

"dog" linear classifier:

$$f\left(x_i; W_{dog}, b_{dog}\right) \quad \text{or}$$

$$f(x_i; W, b)_{dog} \quad \text{or} \quad f_{dog}$$

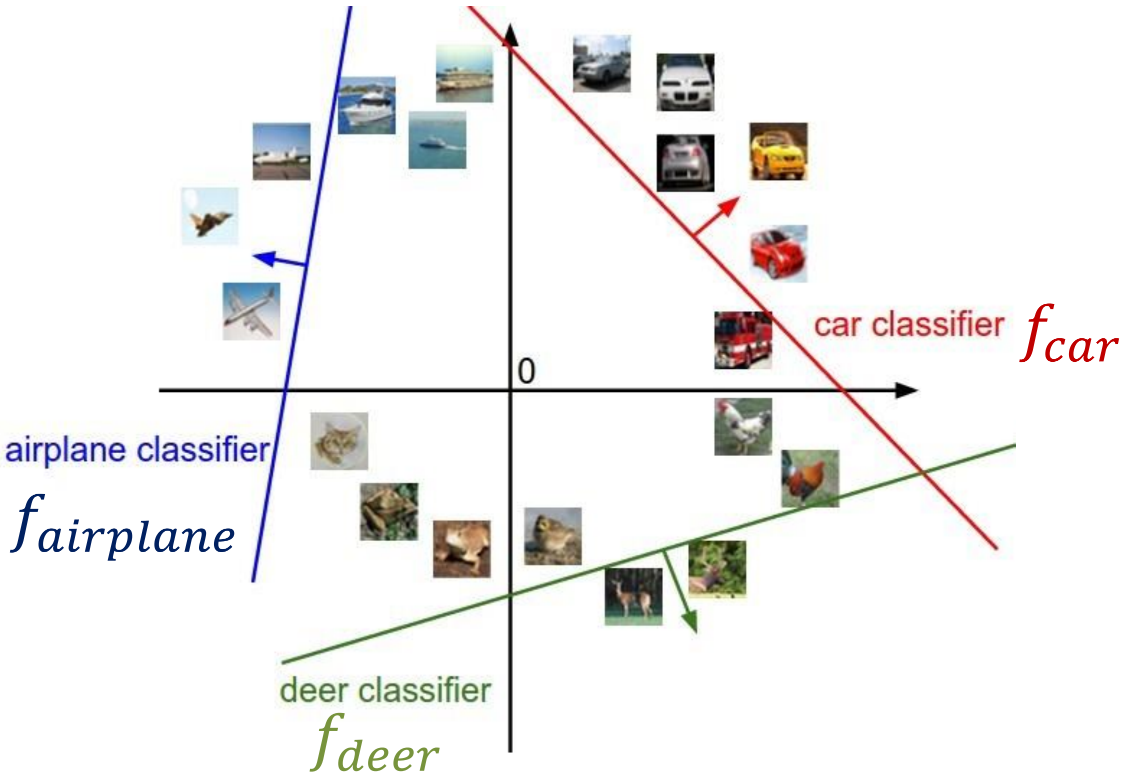
Linear classifier for label *j*:

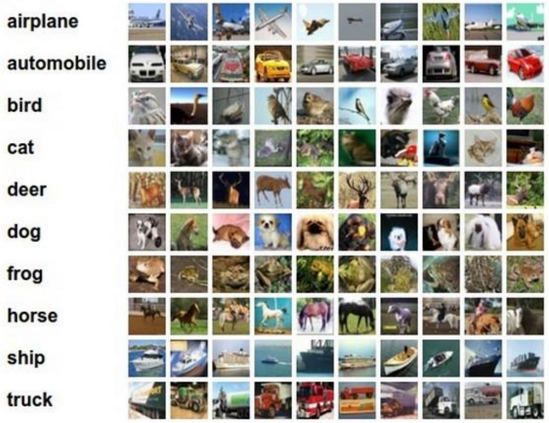
$$f\left(x_i; W_j, b_j\right) \quad \text{or}$$

$$f(x_i; W, b)_j \quad \text{or} \quad f_j$$

Interpreting Multiple Linear Classifiers

$$f(x_i; W_j, b_j) = W_j x_i + b_j$$





CIFAR-10 object recognition dataset

Linear Classification: 2) Loss Function

(or cost function or objective)

Scores $f(x_i; W)$ | **Label** $y_i = 2\ (dog)$ → **Loss** $L_i = ?$

Image x_i

(Size: 32*32*3)

0 (duck) ?	-12.3
1 (cat) ?	45.6
2 (dog) ?	98.7
3 (pig) ?	12.2
4 (bird) ?	-45.3

Multi-class problem

How to assign only one number representing how "unhappy" we are about these scores?

The loss function quantifies the amount by which the prediction scores deviate from the actual values.

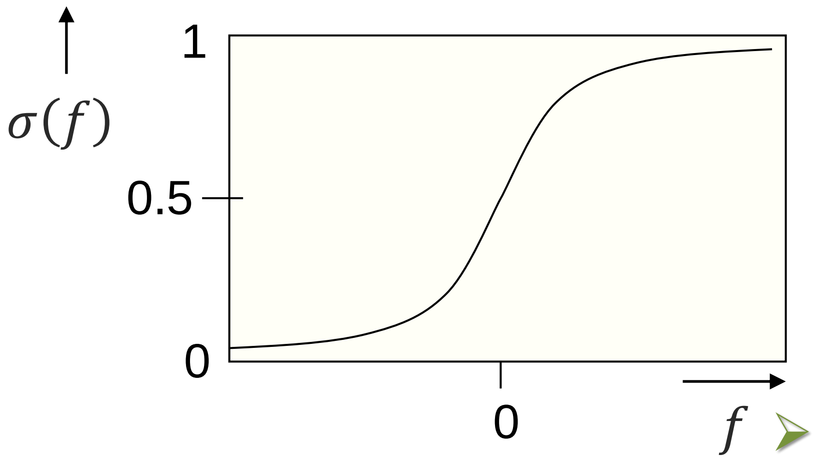
A first challenge: how to normalize the scores?

First Loss Function: Cross-Entropy Loss

(or logistic loss)

Logistic function:

$$\sigma(f) = \frac{1}{1 + e^{-f}}$$



➢ **Score function**

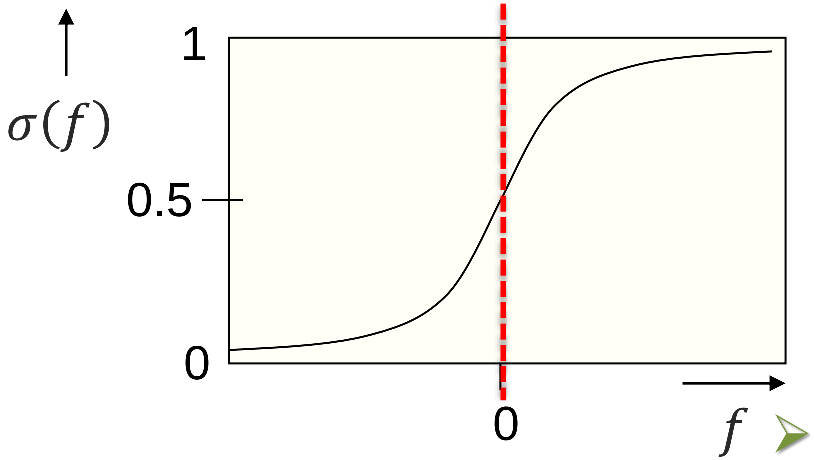
First Loss Function: Cross-Entropy Loss

(or logistic loss)

Logistic function: $$\sigma(f) = \frac{1}{1 + e^{-f}}$$

Logistic regression: (two classes) $$p(y_i = "dog" | x_i; w) = \sigma(w^T x_i)$$

= true for two-class problem



➢ **Score function**

First Loss Function: Cross-Entropy Loss

(or logistic loss)

Logistic function:

$$\sigma(f) = \frac{1}{1 + e^{-f}}$$

Logistic regression:
(two classes)

$$p(y_i = "dog" | x_i; w) \quad = \sigma(w^T x_i)$$

= true
for two-class problem

Softmax function:
(multiple classes)

$$p(y_i | x_i; W) = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}}$$

First Loss Function: Cross-Entropy Loss

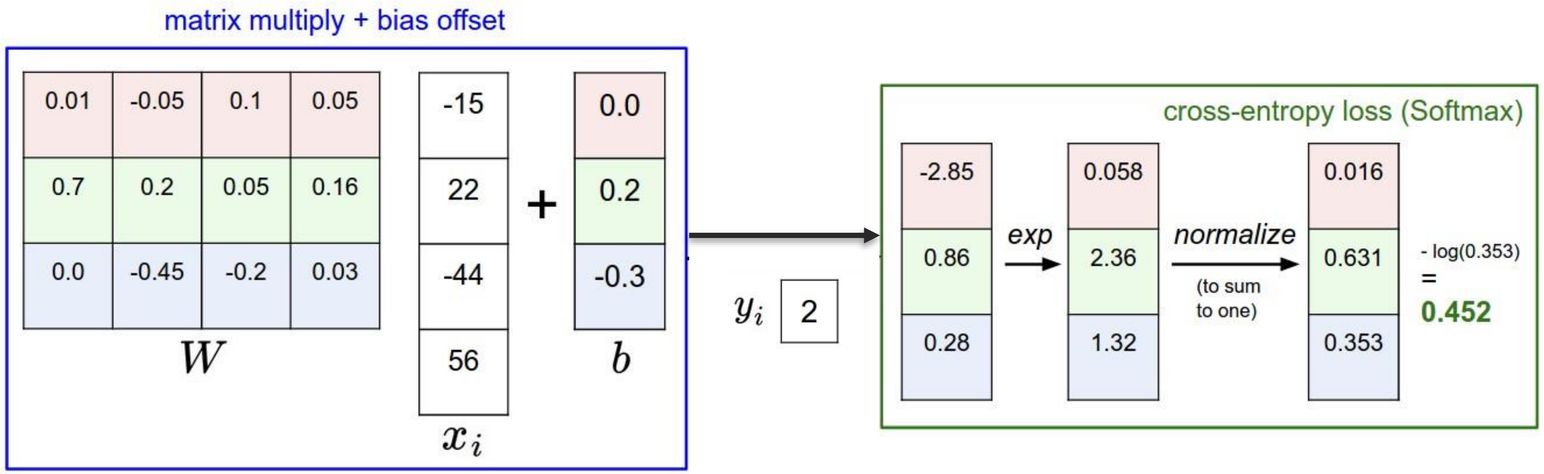
(or logistic loss)

Cross-entropy loss:

$$L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$$

Softmax function

Minimizing the negative log likelihood.



Second Loss Function: Hinge Loss

(or max-margin loss or Multi-class SVM loss)

$$L_i = \sum_{j \neq y_i} \max(0, f(x_i, W)_j - f(x_i, W)_{y_i} + \Delta)$$

loss due to example i

sum over all incorrect labels

difference between the correct class score and incorrect class score

Second Loss Function: Hinge Loss

(or max-margin loss or Multi-class SVM loss)

$$L_i = \sum_{j \neq y_i} \max(0, f(x_i, W)_j - f(x_i, W)_{y_i} + \Delta)$$

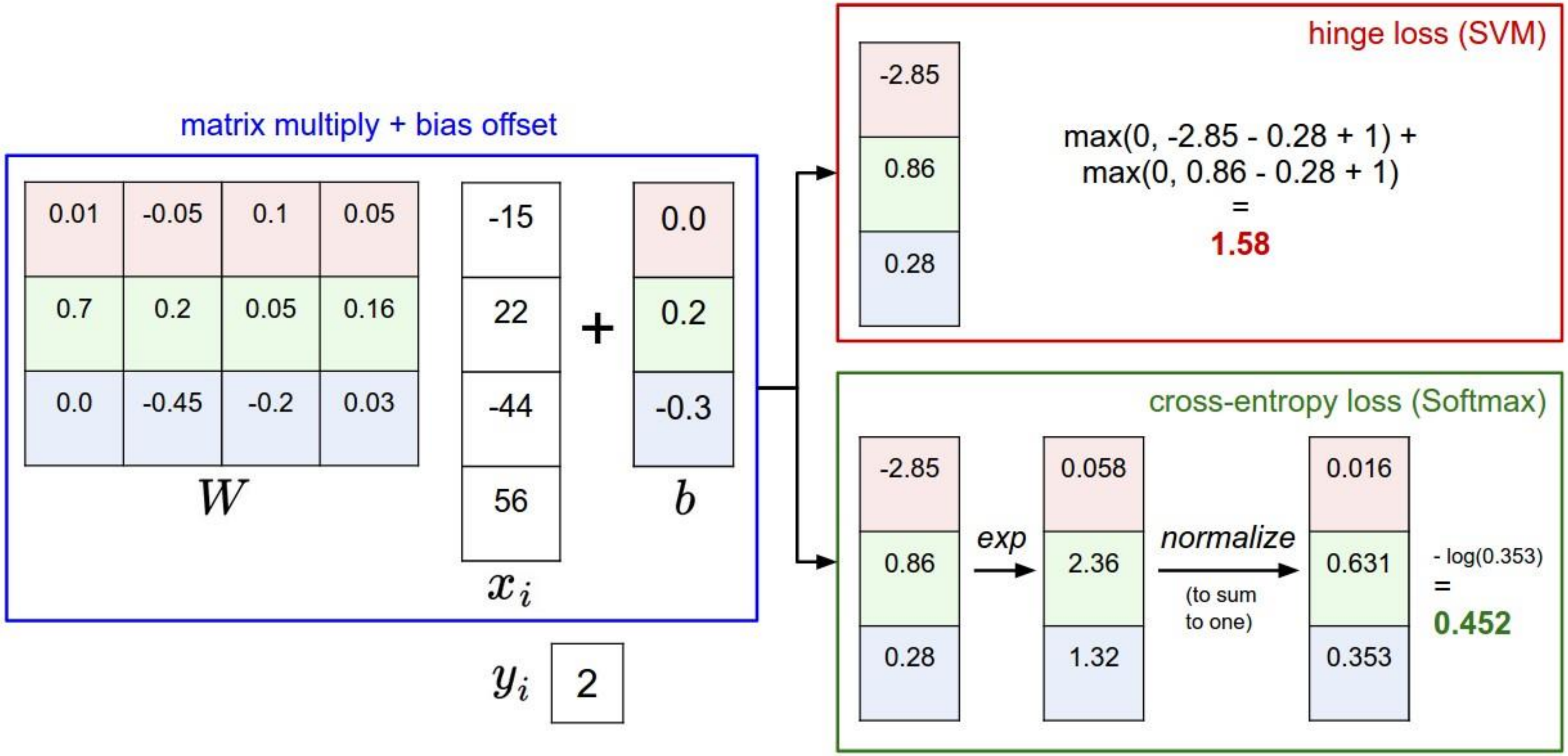
e.g. 10

Example: $f(x_i, W) = [13, -7, 11]$

$y_i = 0$

$$L_i = \max(0, -7 - 13 + 10) + \max(0, 11 - 13 + 10)$$

Two Loss Functions



Loss Function

Loss function is often made up of three parts

$$L = L_{data} + \lambda_1 L_{regularization} + \lambda_2 L_{constraints}$$

1. Data term

   How well our model is explaining/predicting training data

   e.g. cross-entropy loss, Euclidean loss

$$\sum_i L_i = -\sum_i \log\left(\frac{e^{f_{y_i}(x_i;W)}}{\sum_j e^{f_j(x_i;W)}}\right)$$

$$\sum_i L_i = \sum_i \left(y_i - f(x_i, W)\right)^2$$

Loss Function

Loss function is often made up of three parts

$$L = L_{data} + \lambda_1 L_{regularization} + \lambda_2 L_{constraints}$$

2. Regularization/Smoothness term

   Prevent the model from becoming too complex

      e.g. $||W||_2$ for parameters smoothness

      e.g. $||W||_1$ for parameter sparsity

   λ_1 is a hyper-parameter

      Optional, but almost never omitted

Loss Function

Loss function is often made up of three parts

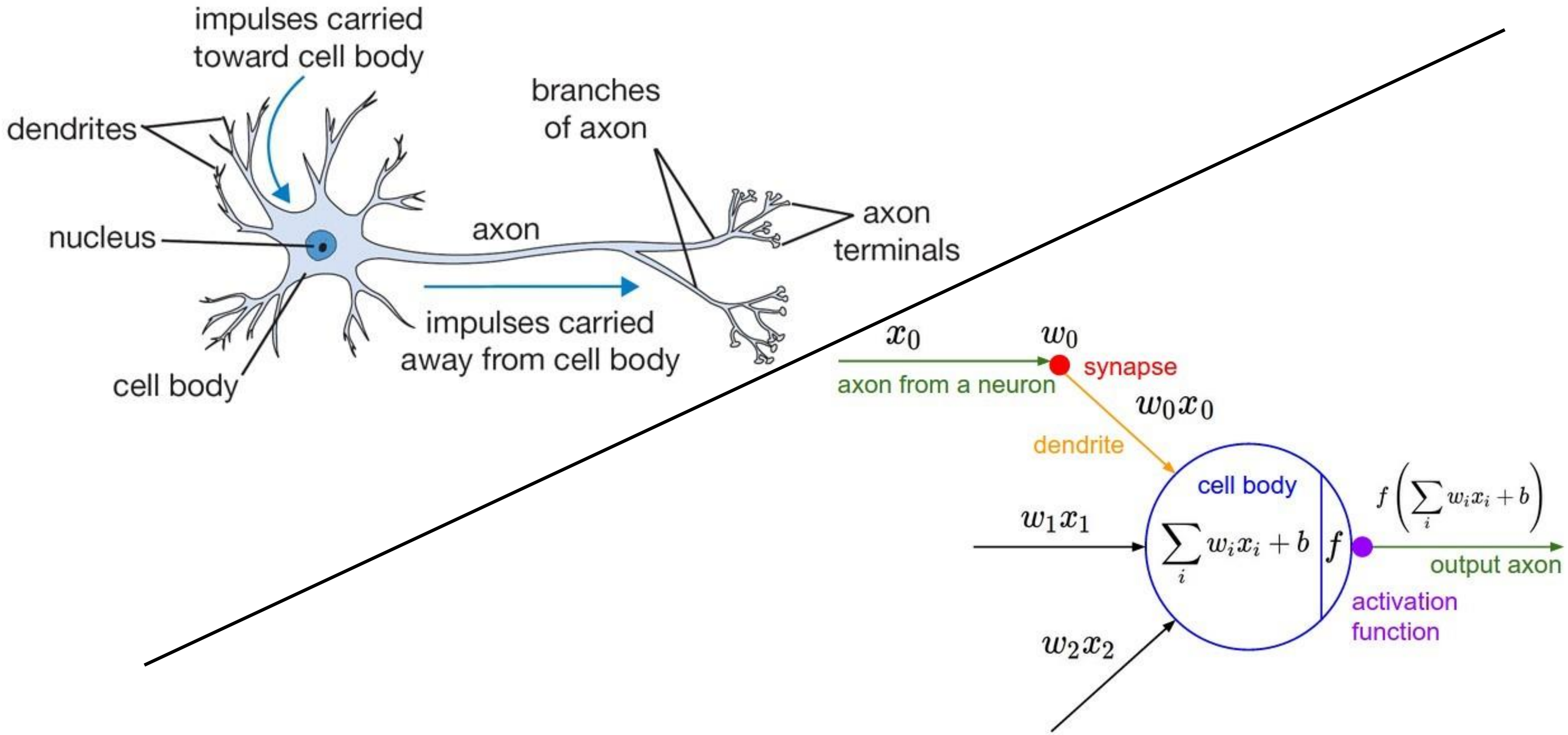
$$L = L_{data} + \lambda_1 L_{regularization} + \lambda_2 L_{constraints}$$

3. Additional constraints
   - Optional and not always used. Help with certain models
     - Example during lecture 3.2 about coordinated multimodal representation
   - Example of loss functions using constraints:
     - Triplet loss, hinge ranking loss, reconstruction loss

Basic Concepts: Neural Networks

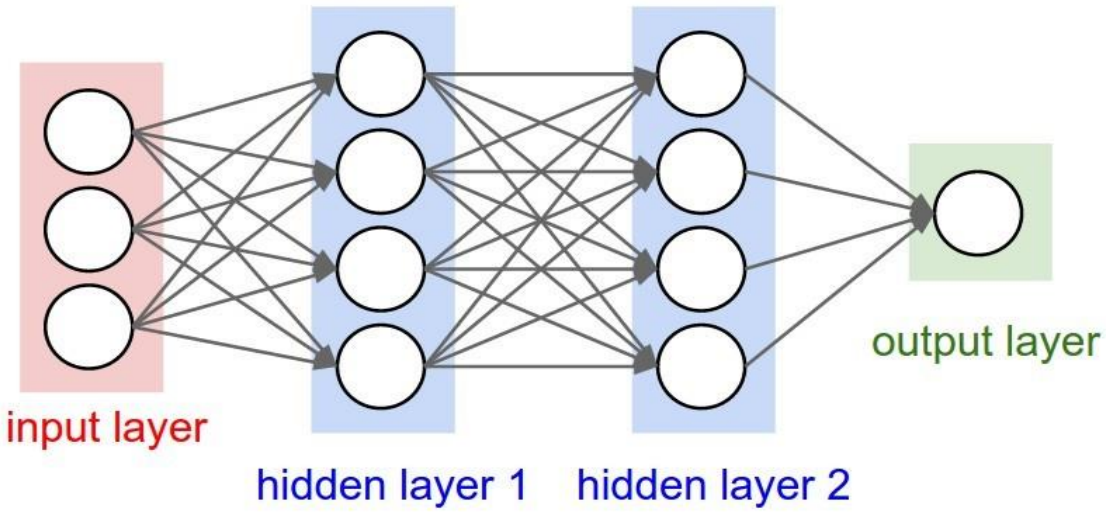
Neural Networks – inspiration

- Made up of artificial neurons



Neural Networks – score function

- Made up of artificial neurons
  - Linear function (dot product) followed by a nonlinear activation function
- Example a Multi Layer Perceptron



Basic NN building block

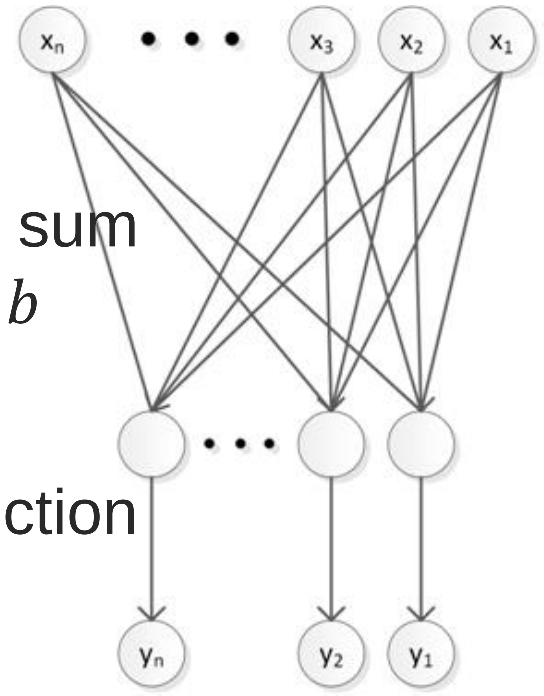
- Weighted sum followed by an activation function

Input

Weighted sum
$Wx + b$

Activation function

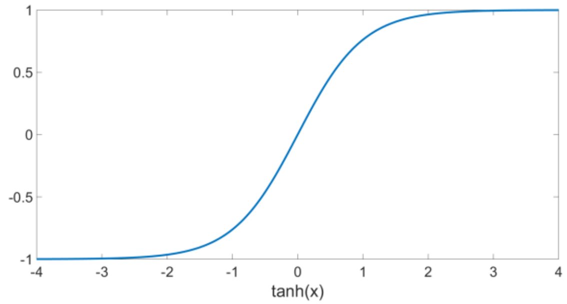
Output



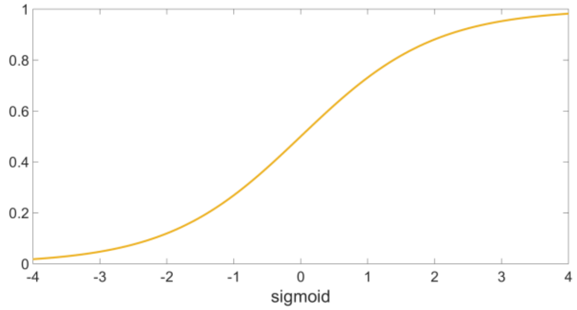
$$y = f(Wx + b)$$

Neural Networks – activation function

- $f(x) = \tanh(x)$

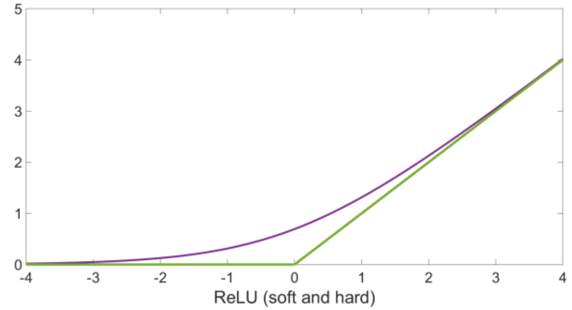


- Sigmoid - $f(x) = (1 + e^{-x})^{-1}$



- Linear – $f(x) = ax + b$

- ReLU $f(x) = \max(0, x) \sim \log(1 + \exp(x))$
  - Rectifier Linear Units
  - Faster training - no gradient vanishing
  - Induces sparsity



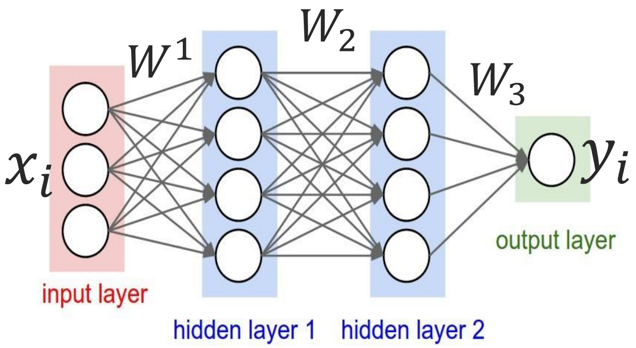
Multi-Layer Feedforward Network

Activation functions (individual layers)

$$f_{1;W_1}(x) = \sigma(W_1 x + b_1)$$

$$f_{2;W_2}(x) = \sigma(W_2 x + b_2)$$

$$f_{3;W_3}(x) = \sigma(W_3 x + b_3)$$



Score function

$$y_i = f(x_i) = f_{3;W_3}(f_{2;W_2}(f_{1;W_1}(x_i)))$$

Loss function (e.g., Euclidean loss)

$$L_i = (f(x_i) - y_i)^2 = (f_{3;W_3}(f_{2;W_2}(f_{1;W_1}(x_i))))^2$$

Neural Networks inference and learning

- Inference (Testing)
  - Use the score function ($y = f(\boldsymbol{x}; W)$)
  - Have a trained model (parameters W)
- Learning model parameters (Training)
  - Loss function (L)
  - Gradient
  - Optimization

Don’t Forget! Course Assignments…

Today 8pm: Project preference form

Tomorrow 8pm: Your reading selection
(using the Google Sheet for your study group)

Friday 8pm: Post your summary

Monday 8pm: Follow-up posts about other papers